MediaEval 2016: A multimodal system for the Verifying Multimedia Use task

Cédric Maigrot[1], Vincent Claveau[2], Ewa Kijak[1], and Ronan Sicre[2]

[1,2] IRISA, [1]Univ. of Rennes 1, [2]CNRS, Rennes, France ,

Cedric.Maigrot@irisa.fr, *Vincent.Claveau@irisa.fr*, *Ewa.Kijak@irisa.fr*, *Ronan.Sicre@irisa.fr*

ABSTRACT

This paper presents a multi-modal hoax detection system composed of text, source, and image analysis. As hoax can be very diverse, we want to analyze several modalities to better detect them. This system is applied in the context of the *Verifying Multimedia Use* task of MediaEval 2016. Experiments show the performance of each separated modality as well as their combination.

1. INTRODUCTION

Social Networks (SN) have been of increasing importance in our daily lives. When studying SN, one interesting aspect is the publication propagation, *e.g.* news, facts, or any information considered as important and shared across communities. A major characteristic of the propagation is its speed. However, users rarely verify the veracity of the shared information. Moreover, verified false information is often shared and spreading can not be contained [11, 9].

Therefore, we are studying how to verify directly the veracity of any information. Our goal is to create systems that can inform users before sharing false information. Consequently, we are extremely interested in the *Verification Multimedia Use* task of *MediaEval 2016*, which aims at classifying Twitter publications to detect fake information [2]. Considering the nature of tweet data, diverse information coming from the message and its meta-data can be extracted. We explored in this work the predictive power of various features. We propose different approaches based on text information, source credibility, and image content.

2. APPROACHES

We propose four approaches: text-based (`run-T`), source-based (`run-S`), image-based (`run-I`), and the combination of the three approaches (`run-C`). For all of these methods the prediction is first made at the image-level, then propagated to the tweets that contains the image, according to the following rule: the tweet is predicted as *real* if all the associated images are classified as *real*; if at least one of the images is classified as *fake*, the tweet is considered as *fake*.

2.1 Text-based nearest neighbors prediction

This approach exploits the textual contents of the tweets and do not rely on any external data apart from the training set. As previously explained, a tweet is classified based on the images it contains; an image is described by the concatenated texts of every tweet containing this image.The idea here is to capture similar comments between an unknown image and an image from the training set (such as *It's photoshopped*) or similar genres of comments (presence of smileys, slang/journalistic languages...).

Let us note I_q such a description for an unknown image, and $\{I_{d_i}\}$ the training set of image descriptions. The class of I_q is decided based on the classes of the k similar image descriptions in $\{I_{d_i}\}$. In practice, to compute the similarities, we use a state-of-the-art information retrieval approach called Okapi-BM25 [5]. A language-detection system (based on the Google translate service[1]) is used to detect non English tweets, which are then translated into English with Google translate. As another preprocessing, we use orthographic and smiley normalization tools developed in-house. The parameter k was set to 1 by cross-validation.

2.2 Trusted sources prediction

This approach, already used by [4], is conceptually the simplest but rely on external (static) knowledge. As for the previous run, prediction is made at the image level, and an image is represented as the concatenation of every tweet (translated in English if needed) in which it appears. The prediction is made by detecting *trustworthy* sources in the image description. Two types of sources are searched: 1) a known news-related organism; 2) an explicit citation of the source of the image. For the first types, we gathered lists of press agencies in the world, newspapers (mostly French and English ones), news TV networks (French and English ones). For the second types, we manually defined some patterns, like `photographed by + Name`, `captured by + Name`, etc. Finally, an image is classified as fake by default, unless a trustworthy source is found in its text description.

2.3 Image retrieval prediction

In this approach only the image content is used to provide a prediction, at the image level. Note that some tweets do not contain images but videos; such tweets are thus labeled as *unknown*.

[1]https://translate.google.com/

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

Images from the *Verification Multimedia Use* task are classified using external information. We perform image retrieval, which consists in querying a database of known fake/real images to discover already known fake images. The database is built by collecting images from 5 specialized websites, *i.e www.hoaxbuster.com/*, *hoax-busters.org*, *urbanlegends.about.com*, *snopes.com*, and *www.hoax-slayer.com/*. The set contains around 500 original images and 7500 fake samples.

Generic image descriptors are computed using the very deep Convolutional Neural Networks (CNN) [8]. First, we apply the convolutional layers [10] of the network on images scaled to a standard size of 544×544. Then, the two first fully connected layers are kernelized and applied, on the output feature map, producing a new $11 \times 11 \times 4096$ dimensional feature map. Finally, average pooling followed by *l2*-normalization is performed, giving a 4096-dimensional descriptor [3, 7, 6]. Once all images descriptors are obtained, cosine similarity is computed between the query and all images from the database. If the highest similarity is higher than a threshold of 0.9 (set on the training dataset), then the query receives the label of the most similar image. Otherwise, the query is labeled as *unknown*.

2.4 Combination

This last approach aims at combining the three preceding ones in a late fusion process. Thus, for a given image, it takes as input the predictions given by the three systems describe above. As before, the final prediction on the image is then propagated to the tweets containing it.

Instead of using a simple fusion process (for instance, a majority vote), we try to automatically build a fusion model fine-tuned to the task. We thus use a machine learning algorithm, namely boosting (adaboost.MH) over decision trees [1], which takes as input the predictions of the three previous approaches, and also the scores associated to these predictions (for run-T and run-I). The parameters of the machine learning algorithm are set by cross-validation on the training data: the number of iterations for boosting is 500 and the depth of the trees is 3. Finally, the fusion model is learned on the whole training set; it is then used on the test set images.

3. RESULTS

The four approaches are applied on the MediaEval 2016 test set and results are reported in Figure 1. The test set is composed of 2228 Twitter messages associated with 130 images. Moreover, 65% and 26% of the tweets of the development and test set respectively are associated with a single event.

We observe that the approach based on the source trustworthiness level (run-S) outperforms the text-based approach (run-T), which outperforms the image-based approach (run-I). We can see that the text-based approach competes with the source-based approach in terms of recall. It means that the text approach tends to classify every tweet as fake. This may be explained by the fact that the training set is unbalanced as it contains 3 times more *fake* than *real*.

We note that the prediction based on the image approach has several drawbacks and performs poorly. In particular, the precision is low compared to what we estimated on the training set. Several explanations can be given. First, only 86% of the test tweets are associated with one or more images (the rest are associated with video content), meaning that the image approach is evaluated only on this portion of the dataset. Therefore, recall and F-score are directly impacted. Secondly, the reference database that we built is small and unbalanced, resulting in a high number of *unknown* labels in the predictions. Thirdly, the base does not always contain the original images and small modifications between forged image and its original version can be considered as similar. Finally, images shared on SN often present specific editing characteristics, as visible added watermarks like *fake*,*rumor* or *real*, circles, text annotations, etc. Such edits impair the similarity computation between images.

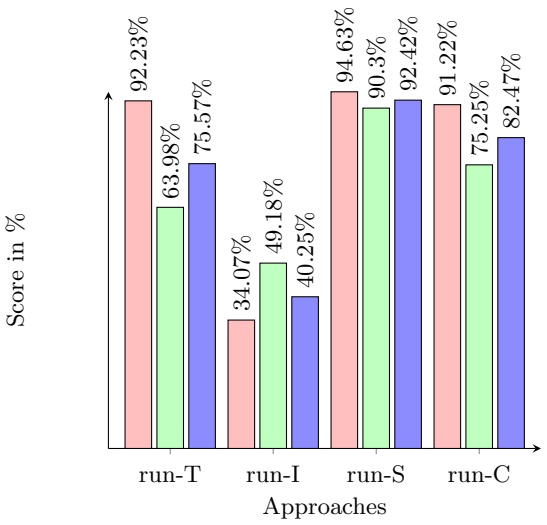

Figure 1: Recall (red), precision (green) and F-Measure (blue) scores of the *fake* class on the test set.

Concerning the run-C, we note that the combination using late fusion does not offer any gain, and perform even worse than the run-S alone. This result is disappointing, as it differs from what we evaluated on the training set by cross-validation. It may be explained by an overfitting problem when learning the fusion model, and by the lower precision (compared to the one estimated on training set) obtained by the run-I which is used as input.

4. CONCLUSION

A multi-modal hoax detection system based on text, source, and image analysis is presented. This system uses different categories of external knowledge: static and general ones, such as press agency lists, and dynamic and dedicated ones such as hoax listing websites, etc. Our evaluation conforts previous results on the good performance of the source analysis; conversely, the image approach shows poor results. Yet, we still consider this later approach as promising; several improvements are foreseen to improve both the database and the content comparison. Finally, multimodality remains a challenge, as integrating different sources of knowledge may result in performance loss.

5. ACKNOWLEDGEMENTS

This work is partly supported by the Direction Générale de l'Armement, France (DGA).

References

[1] Nathalie Camelin Antoine Laurent and Christian Raymond. Boosting bonsai trees for efficient features combination : application to speaker role identification. In *Proc. of InterSpeech*, 2014.

[2] Christina Boididou, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, Michael Riegler, Stuart E. Middleton, Katerina Andreadou, and Yiannis Kompatsiaris. Verifying multimedia use at mediaeval 2016. In *Working Notes Proceedings of the MediaEval 2016 Workshop*, 2016.

[3] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, and Andrea Vedaldi. Deep filter banks for texture recognition, description, and segmentation. *International Journal of Computer Vision*, 118(1):65–94, 2016.

[4] Stuart Middleton. Extracting attributed verification and debunking reports from social media: mediaeval-2015 trust and credibility analysis of image and video. 2015.

[5] Stephen E. Robertson, Steve Walker, and Micheline Hancock-Beaulieu. Okapi at TREC-7: Automatic Ad Hoc, Filtering, VLC and Interactive. In *Proc. of the 7^{th} Text Retrieval Conference, TREC-7*, pages 199–210, 1998.

[6] Ronan Sicre and Hervé Jégou. Memory vectors for particular object retrieval with multiple queries. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 479–482. ACM, 2015.

[7] Ronan Sicre and Frédéric Jurie. Discriminative part model for visual recognition. *Computer Vision and Image Understanding*, 141:28–37, 2015.

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[9] Hokky Situngkir. Spread of hoax in social media. 2011.

[10] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. Particular object retrieval with integral max-pooling of cnn activations. In *ICLR*, 2016.

[11] Jaewon Yang and Jure Leskovec. Modeling information diffusion in implicit networks. In *2010 IEEE International Conference on Data Mining*, pages 599–608. IEEE, 2010.